DOCUMENT RESUME

ED 316 583 TM 014 558

AUTHOR Welge, Patricia
TITLE Three Reasons Why Stepwise Regression Methods Should Not Be Used by Researchers.
PUB DATE Jan 90
NOTE 22p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Austin, TX, January 25-27, 1990).
PUB TYPE Information Analyses (070) -- Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Computer Software; Doctoral Dissertations; *Predictor Variables; *Regression (Statistics); Researchers; Research Methodology; *Research Problems; Social Science Research
IDENTIFIERS *Stepwise Regression; Variance (Statistical)

ABSTRACT

L. A. Marascuilo and R. C. Serlin (1988) note that stepwise regression is a method used frequently in social science research. C. Huberty (1989) characterizes such applications as being "common". In support of this latter statement, a review of dissertations by B. Thompson (1988) demonstrated that dissertation students frequently use stepwise methods in their data analyses. Unfortunately, stepwise regression can lead to serious misinterpretations of results. Three fundamental problems with traditional stepwise regression analysis are discussed in this paper. Firstly, most researchers who use "canned" computer programs do not use the correct degrees of freedom when evaluating changes in explained variance. Secondly, there is a tendency for some researchers to incorrectly interpret order of selection and entry of the "g" predictor variables as defining the "best" subset of predictors of size "g". Thirdly, some researchers incorrectly consult order of entry information to evaluate the importance of various predictor variables. A small data set is used to illustrate the difficulties with stepwise methods, and it is argued that stepwise methods should not be used by researchers when the desired outcome is accurate variable selection or ordering of variables with respect to importance. Three data tables are included. (Author/TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED316583

stepwise.rno 1/8/90

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

PATRICIA WELGE

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Three Reasons Why Stepwise Regression Methods Should Not Be Used By Researchers

Patricia Welge

School of Allied Health Professions

Louisiana State University Medical Center

Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX, January 27, 1990.

TM014558

BEST COPY AVAILABLE

Abstract

Marascuilo and Serlin (1988, p. 671) note that stepwise regression is a method used frequently in social science research. Huberty (1989) characterizes such applications as being "common". In support of this statement, a review of dissertations by Thompson (1988) demonstrated that dissertation students frequently employed stepwise methods in their data analysis. Unfortunately, stepwise regression can lead to serious misinterpretations of results. Three fundamental problems with traditional stepwise regression analysis are discussed in this paper. A small data set is used to illustrate the difficulties with stepwise methods, and it is argued that stepwise methods should not be used by researchers when the desired outcome is accurate variable selection or ordering of variables with respect to importance.

Three Reasons Why Stepwise Regression Methods Should Not Be Used By Researchers

Stepwise regression methods are viewed as a variation of forward solution methods employed in some multiple regression analyses (Cliff, 1987; Huberty, 1989). In the forward solution method, variables are entered into the regression equation one at a time. The first variable selected is the predictor that has the highest single correlation, in absolute value, with the criterion variable. The next predictor selected is the one which has the highest partial or part correlation with the dependent variable, after the effects of the first variable have been partitioned out. This entry procedure results in the greatest increase in R^2 as each new predictor in entered into the analysis (Cliff, 1987; Hinkle, Wiersma & Jurs, 1988). Subsequent predictor variables are selected in the manner described above at each step until the increase in R^2 is no longer statistically significant, all predictors have been entered, or the researcher is satisfied with the resulting effect size.

Stepwise regression can be viewed as a variation of the forward selection method since predictor variables are entered one at a time, but true stepwise entry differs from forward entry in that at each step of a

stepwise analysis the removal of each entered predictor is also considered; entered predictors are deleted in subsequent steps if they no longer contribute appreciable unique predictive power to the regression when considered in combination with newly entered predictors (Thompson, 1989).

Stepwise analytic methods may be among the most popular analyses employed in both substantive and validity research. Stepwise methods can be employed in both hierarchical and variable-selection multiple regression analyses (Cliff, 1987). Most commonly, however, stepwise methods are employed in variable-selection analyses. As Marascuilo and Serlin (1988, p. 671) noted, "The most common method for selecting the fewest number of predictor variables necessary to guarantee adequate prediction is based on a model referred to as stepwise regression." Huberty (1989) concurred and stated that stepwise analyses have basically been used for three purposes: (a) selection or deletion of variables; (b) assessing relative variable importance; or (c) both variable selection and variable ordering. Huberty pointed out, however, that one should, "Note that these are uses often made of stepwise analyses, not uses that necessarily **should** be made" (p. 45). Stepwise methods were designed primarily to assist the researcher in selecting, from a large pool of

variables, a smaller set of variables which would be as efficient, or almost as efficient, in developing prediction estimates. Thompson (1988) noted, however, that utilization of stepwise methods can lead to serious misinterpretation of results.

Problems with Stepwise Methods

Three major problems with stepwise methods have been identified (cf. Thompson, 1989). First, most researchers who utilize "canned" computer programs do not utilize the correct degrees of freedom when evaluating changes in explained variance. For example, in the stepwise regression analyses of SPSS-X the number of predictor variables (n_{pv}) or degrees of freedom explained at step one is recorded as one, even though results for all predictor variables present in the entire analysis are actually consulted prior to entering the one predictor variable with the largest single-order correlation with the criterion variable.

Similarly, at step two, a second predictor can be added into the equation. The program tests the statistical significance of the change in R^2 by utilizing an F test with 2 and $n-g-1$ degrees of freedom, respectively, where g is the number of predictor variables entered at a given step. The numerator degrees of freedom presumes illogically that only one additional predictor variable was employed to yield the change in

R^2 when, in fact, the results associated with all the predictors were consulted to select the new predictor and thus yield the tested R^2. As Cliff (1987, p. 187) stated, "in a sense all the variables are in the equation, even though some of them have (effectively) been given zero weights." For example, if the 12 person data set in Table 1 were utilized and the nine person sample consisted of all ID numbers except 1, 3, and 7, the calculated F for the nine person sample at the first step, utilizing the liberal degrees of freedom offered by the computer program of 1 and 7 respectively, would be 2.14795, statistically significant at alpha = .25. In contrast, if a more conservative approach were adopted and all four predictor variables were considered in the degrees of freedom numerator, even though some of them have zero weights, the calculated F becomes .3068501, not statistically significant at alpha = .25. These data illustrate why Cliff (1987, p. 185) stated that, "most computer programs for (stepwise) multiple regression are positively satanic in their temptation toward Type I errors."

Insert Table 1 about here

A second problem with stepwise regression methods is the tendency of some researchers to incorrectly interpret order of selection and entry of the g predictor

variables as defining the "best" subset of predictors of size q. Thompson (1989) suggested that in a stepwise analysis in which three steps are conducted, and predictors A, B, and C are selected from among predictors A, B, C, D, E, and F, it is entirely possible that three entirely different predictors (D, E, and F) might be the optimal predictor set of size three. Stepwise selects the next-best predictor at each step, given the presence of the previous predictors. Selecting the next-best predictor for each of q steps is not the same as selecting the optimal predictor variable set of size q. As Huberty (1989) noted:

> It is generally understood by methodologists that the first q variables entered into either a regression analysis or a discriminant analysis do not necessarily constitute the "best" subset of size q... In regression analysis there may be one or more subsets of size q that would, for the data on hand, yield a larger R squared value.... Furthermore, and perhaps more importantly, if a given subset of size q yielded the largest R squared of all subsets of that size, there may very well be other subsets of size q that yield nearly the same R squared value. And one or more of these subsets may be more interesting or relevant in a substantive

sense. (p. 46)

Thompson (1988) cited an example from a dissertation in which the student researcher incorrectly associated order of variable entry with selection of the "best" subset of predictors of size g. This student utilized stepwise regression methods and concluded, based on the stepwise results, that computer experience ($r = -0.288$) and mathematics anxiety ($r = -0.141$) contributed significantly to the variance of the computer anxiety dependent variable. The importance of these two predictors was emphasized. Yet, a third predictor, number of previous math classes, apparently had a larger r (-0.175) with the dependent variable than math anxiety, but was not considered by the student since this third predictor was not part of the stepwise regression entering only two predictors.

A third reason why stepwise regression methods should not be used is that some researchers incorrectly consult order of entry information to evaluate the importance of various predictor variables (Thompson, 1989). As Huberty (1989) explained:

> The first variable entered with a stepwise regression analysis is determined by the correlation between each predictor variable and the criterion variable... The third, say, variable to be entered (and often considered to

be the third most important) is dependent on the two variables already entered. If one or two of the variables already entered would be changed, then the third variable entered may also be different. This dependence or conditionality truly makes variable importance as determined by stepwise analysis **very** questionable. (pp. 46-47)

Sampling error can drastically influence the results of the stepwise regression analysis. The small data set for a population (N = 12) presented in Table 1 is utilized to illustrate how seriously sampling error can distort the interpretation of stepwise results involved in predicting the criterion variable ZY. Table 2 indicates that the three predictor variables share little variance with each other and that the order of the predictor variables with respect to correct order of stepwise entry in the known population is ZX1, ZX2, ZX3, and ZX4.

Insert Table 2 about here

Presume that the researcher drew a random sample of nine subjects from the population. Each of the 220 combinations of nine participants (i.e., omit participants 1, 2, 3; omit 1, 2, 4, and so on) would be exactly equally probable. As reported in Table 3, for these illustrative data, only 35 of the 220 equally

probable possible samples entered the predictor variables in the order that is correct in the full, known population parameters. Only 87 samples of the 220 possible samples selected ZX1 as the first prediction entry, as this variable is in the full population. Sixty-eight samples selected ZX2 as the first entry; 40 samples selected ZX3 as the first variable entered; and 25 samples selected the worst predictor, ZX4, as the best predictor of ZY.

Insert Table 3 about here

It is clear from this small example data set that order of variable selection in a sample does not give the researcher stable or generalizable information regarding the actual importance of various predictors. As Huberty (1989) states:

> Inferences about "best" subsets and variable importance should be made with great caution. The "best" variable subset for one sample of units may be far from the best for other samples. The greater the ratio of sample size to number of response variables, the more reasonable are the implied generalizations. A large ratio alone, however, does not insure valid generalizations. Valid generalizations may be obtained only to the extent that the

> pattern of response variable intercorrelations for non-design sample experimental units follow the pattern present in the design sample. (p. 63)

If a researcher wishes to draw conclusions regarding variable selection and ordering based on obtained sample results when stepwise methods are employed, then one should always, at a minimum, employ a replication procedure in a different sample to determine how well the obtained regression equation predicts values on the dependent variable for other subjects (Kachigan, 1986). Alternatively, as Thompson (1989) and Huberty (1989) have suggested, the researcher could utilize a cross-validation procedure, or a re-sampling strategy such as the jackknife or the bootstrap. As Kachigan (1986) argued, these validation procedures are essential if the researcher wishes to have any faith at all in the generalizability of the obtained regression results.

Conclusions

According to Huberty (1989), obtaining a final, definitive analysis regarding variable selection and ordering with one run via a computer program package that employs methods such as stepwise regression will simply not do the job for most empirical investigations. Similarly, Kerlinger (1986) argues that "the research problem and the theory behind the problem (and not

stepwise methods) should determine the order of entry of variables in multiple regression analysis" (p. 545). Huberty (1989) concurred, suggesting that, "Data should be carefully examined, using multiple analyses and good sound judgment" (p. 64). Thus, researchers who choose to employ stepwise methods, particularly if they also fail to use replication or cross-validation methods, might best consider Cliff's (1987, pp. 120-121) argument that "a large proportion of the published results using this method probably present conclusions that are not supported by the data."

References

Cliff, N. (1987). Analyzing multivariate data. New York: Harcourt Brace Jovanovich.

Hinkle, D.E., Wiersma, W. & Jurs, S.G. (1988). Applied statistics for the behavioral sciences (2nd ed.). Boston: Houghton Mifflin Co.

Huberty, C. (1989). Problems with stepwise methods--better alternatives. In B. Thompson (Ed.), Advances in social science methodology (Vol. 1) (pp.46-47,63-64). Greenwich, CT: JAI Press.

Kachigan, S. K. (1986). Statistical analysis: An interdisciplinary introduction to univariate and multivariate methods (2nd ed.). New York: Radius Press.

Kerlinger, F.N. (1986). Foundations of behavioral research (3rd ed.). New York: Holt, Rinehart and Winston.

Marascuilo, L.A. & Serlin, R.C. (1988). Statistical methods for the social and behavioral sciences. New York: W.H. Freeman.

Pedhazur, E. (1982). Multiple regression in behavioral research (2nd ed.). New York: Holt, Rinehart and Winston.

Thompson, B. (1988, November). Common methodology mistakes in dissertations: Improving dissertation quality. Paper presented at the annual meeting of

the Mid-South Educational Research Association, Louisville, KY. (ERIC Document Reproduction Service No. ED 301 595)

Thompson, B. (1989). Why won't stepwise methods die? Measurement and Evaluation in Counseling and Development, 21(4), 146-148.

Table 1

Standardized Data for Five Variables

ID	ZY	ZX1	ZX2	ZX3	ZX4
1	.790	1.422	.350	.322	-.313
2	-1.589	.112	-1.239	-1.094	-.365
3	.127	-.965	.271	.201	-.060
4	-1.656	-2.167	-.498	-.970	.218
5	.176	-1.291	.153	2.393	.159
6	-.017	.636	-1.607	-.168	-1.746
7	-.397	-.173	.931	-.112	-1.704
8	-.594	.532	-.108	.092	.127
9	.846	.528	1.237	-.092	.035
10	.810	.642	-1.400	1.135	1.654
11	1.764	.373	1.290	-.543	1.005
12	-.260	.352	.620	-1.163	.989

Note. Adapted from Thompson (1989).

Table 2

Bivariate Correlation Matrix

	ZY	ZX1	ZX2	ZX3	ZX4
ZY	---	.497	.444	.384	.319
ZX1	24.7%	---	.018	-.074	-.004
ZX2	19.7%	.0%	---	-.054	.099
ZX3	14.7%	.5%	.3%	---	.103
ZX4	10.2%	.0%	1.0%	1.1%	---

Note. Bivariate r coefficients are presented above the diagonal. Common variance (r^2) percentages are presented below the diagonal. Adapted from Thompson (1989).

Table 3

Entry Orders for Table 1 Data when Different Combinations (c=220) of 3 out of 12 Subjects (n=12-3=9) are Omitted

				Entry Sequence																							
				1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3	3	4	4	4	4	4	4
				2	2	3	3	4	4	1	1	3	3	4	4	1	1	2	2	4	4	1	1	2	2	3	3
				3	4	4	2	2	3	3	4	4	1	1	3	2	4	4	1	1	2	2	3	3	1	1	2
Case	Omit			4	3	2	4	3	2	4	3	1	4	3	1	4	2	1	4	2	1	3	2	1	3	2	1
1	1	2	3						1																		
2	1	2	4																						1		
3	1	2	5	1																							
4	1	2	6				1																				
5	1	2	7	1																							
6	1	2	8				1																				
7	1	2	9				1																				
8	1	2	10							1																	
9	1	2	11				1																				
10	1	2	12					1																			
11	1	3	4																					1			
12	1	3	5	1																							
13	1	3	6											1													
14	1	3	7							1																	
15	1	3	8				1																				
16	1	3	9				1																				
17	1	3	10							1																	
18	1	3	11												1												
19	1	3	12	1																							
20	1	4	5																					1			
21	1	4	6																					1			
22	1	4	7									1															
23	1	4	8																					1			
24	1	4	9																					1			
25	1	4	10									1															
26	1	4	11												1												
27	1	4	12																						1		
28	1	5	6		1																						
29	1	5	7	1																							
30	1	5	8	1																							
31	1	5	9	1																							
32	1	5	10							1																	
33	1	5	11																1								
34	1	5	12	1																							
35	1	6	7											1													
36	1	6	8				1																				
37	1	6	9																					1			
38	1	6	10											1													
39	1	6	11												1												
40	1	6	12										1														

				Entry Sequence																							
				1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3	3	4	4	4	4	4	4
				2	2	3	3	4	4	1	1	3	3	4	4	1	1	2	2	4	4	1	1	2	2	3	3
				3	4	4	2	2	3	3	4	4	1	1	3	2	4	4	1	1	2	2	3	3	1	1	2
Case	Omit			4	3	2	4	3	2	4	3	1	4	3	1	4	2	1	4	2	1	3	2	1	3	2	1
41	1	7	8	1																							
42	1	7	9							1																	
43	1	7	10											1													
44	1	7	11													1											
45	1	7	12							1																	
46	1	8	9				1																				
47	1	8	10							1																	
48	1	8	11													1											
49	1	8	12	1																							
50	1	9	10							1																	
51	1	9	11													1											
52	1	9	12	1																							
53	1	10	11										1														
54	1	10	12							1																	
55	1	11	12													1											
56	2	3	4																			1					
57	2	3	5					1																			
58	2	3	6				1																				
59	2	3	7	1																							
60	2	3	8						1																		
61	2	3	9						1																		
62	2	3	10	1																							
63	2	3	11				1																				
64	2	3	12					1																			
65	2	4	5																			1					
66	2	4	6																				1				
67	2	4	7							1																	
68	2	4	8																			1					
69	2	4	9																			1					
70	2	4	10							1																	
71	2	4	11																				1				
72	2	4	12																								1
73	2	5	6					1																			
74	2	5	7	1																							
75	2	5	8					1																			
76	2	5	9					1																			
77	2	5	10	1																							
78	2	5	11				1																				
79	2	5	12					1																			
80	2	6	7	1																							
81	2	6	8				1																				
82	2	6	9				1																				
83	2	6	10								1																
84	2	6	11				1																				
85	2	6	12					1																			

				Entry Sequence																							
				1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3	3	4	4	4	4	4	4
				2	2	3	3	4	4	1	1	3	3	4	4	1	1	2	2	4	4	1	1	2	2	3	3
				3	4	4	2	2	3	3	4	4	1	1	3	2	4	4	1	1	2	2	3	3	1	1	2
Case	Omit			4	3	2	4	3	2	4	3	1	4	3	1	4	2	1	4	2	1	3	2	1	3	2	1
86	2	7	8	1																							
87	2	7	9	1																							
88	2	7	10											1													
89	2	7	11				1																				
90	2	7	12	1																							
91	2	8	9					1																			
92	2	8	10	1																							
93	2	8	11				1																				
94	2	8	12					1																			
95	2	9	10	1																							
96	2	9	11			1																					
97	2	9	12					1																			
98	2	10	11				1																				
99	2	10	12		1																						
100	2	11	12				1																				
101	3	4	5	1																							
102	3	4	6												1												
103	3	4	7									1															
104	3	4	8																							1	
105	3	4	9																						1		
106	3	4	10							1																	
107	3	4	11													1											
108	3	4	12																							1	
109	3	5	6		1																						
110	3	5	7	1																							
111	3	5	8	1																							
112	3	5	9	1																							
113	3	5	10							1																	
114	3	5	11													1											
115	3	5	12		1																						
116	3	6	7											1													
117	3	6	8				1																				
118	3	6	9				1																				
119	3	6	10							1																	
120	3	6	11													1											
121	3	6	12		1																						
122	3	7	8	1																							
123	3	7	9	1																							
124	3	7	10											1													
125	3	7	11													1											
126	3	7	12							1																	
127	3	8	9				1																				
128	3	8	10							1																	
129	3	8	11				1																				
130	3	8	12		1																						

				Entry Sequence																							
				1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3	3	4	4	4	4	4	4
				2	2	3	3	4	4	1	1	3	3	4	4	1	1	2	2	4	4	1	1	2	2	3	3
				3	4	4	2	2	3	3	4	4	1	1	3	2	4	4	1	1	2	2	3	3	1	1	2
Case	Omit			4	3	2	4	3	2	4	3	1	4	3	1	4	2	1	4	2	1	3	2	1	3	2	1
131	3	9	10							1																	
132	3	9	11													1											
133	3	9	12	1																							
134	3	10	11				1																				
135	3	10	12							1																	
136	3	11	12													1											
137	4	5	6											1													
138	4	5	7										1														
139	4	5	8																1								
140	4	5	9															1									
141	4	5	10							1																	
142	4	5	11																1								
143	4	5	12																						1		
144	4	6	7											1													
145	4	6	8																					1			
146	4	6	9																					1			
147	4	6	10											1													
148	4	6	11													1											
149	4	6	12																						1		
150	4	7	8										1														
151	4	7	9										1														
152	4	7	10											1													
153	4	7	11													1											
154	4	7	12											1													
155	4	8	9																						1		
156	4	8	10							1																	
157	4	8	11													1											
158	4	8	12																						1		
159	4	9	10							1																	
160	4	9	11													1											
161	4	9	12																						1		
162	4	10	11										1														
163	4	10	12							1																	
164	4	11	12													1											
165	5	6	7									1															
166	5	6	8	1																							
167	5	6	9		1																						
168	5	6	10								1																
169	5	6	11															1									
170	5	6	12		1																						
171	5	7	8	1																							
172	5	7	9	1																							
173	5	7	10											1													
174	5	7	11																1								
175	5	7	12	1																							

				Entry Sequence																							
				1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3	3	4	4	4	4	4	4
				2	2	3	3	4	4	1	1	3	3	4	4	1	1	2	2	4	4	1	1	2	2	3	3
				3	4	4	2	2	3	3	4	4	1	1	3	2	4	4	1	1	2	2	3	3	1	1	2
Case	Omit			4	3	2	4	3	2	4	3	1	4	3	1	4	2	1	4	2	1	3	2	1	3	2	1
176	5	8	9	1																							
177	5	8	10							1																	
178	5	8	11													1											
179	5	8	12		1																						
180	5	9	10							1																	
181	5	9	11													1											
182	5	9	12		1																						
183	5	10	11													1											
184	5	10	12								1																
185	5	11	12																1								
186	6	7	8							1																	
187	6	7	9											1													
188	6	7	10										1														
189	6	7	11													1											
190	6	7	12											1													
191	6	8	9			1																					
192	6	8	10							1																	
193	6	8	11													1											
194	6	8	12		1																						
195	6	9	10							1																	
196	6	9	11													1											
197	6	9	12		1																						
198	6	10	11										1														
199	6	10	12											1													
200	6	11	12													1											
201	7	8	9	1																							
202	7	8	10											1													
203	7	8	11													1											
204	7	8	12	1																							
205	7	9	10											1													
206	7	9	11													1											
207	7	9	12							1																	
208	7	10	11												1												
209	7	10	12											1													
210	7	11	12													1											
211	8	9	10							1																	
212	8	9	11													1											
213	8	9	12		1																						
214	8	10	11							1																	
215	8	10	12							1																	

| | | | | Entry Sequence | | | |